arXiv:2508.05250v1 [eess.AS] 7 Aug 2025

Privacy Disclosure of Similarity Rank in Speech and Language Processing

Tom Bäckström, *Senior Member, IEEE,* Mohammad Hassan Vali, *Member, IEEE*, My Nguyen, Silas Rech

***Abstract*—Speaker, author, and other biometric identification applications often compare a sample's similarity to a database of templates to determine the identity. Given that data may be noisy and similarity measures can be inaccurate, such a comparison may not reliably identify the true identity as the most similar. Still, even the similarity rank based on an inaccurate similarity measure can disclose private information about the true identity. We propose a methodology for quantifying the privacy disclosure of such a similarity rank by estimating its probability distribution. It is based on determining the histogram of the similarity rank of the true speaker, or when data is scarce, modeling the histogram with the beta-binomial distribution. We express the disclosure in terms of entropy (bits), such that the disclosure from independent features are additive. Our experiments demonstrate that all tested speaker and author characterizations contain personally identifying information (PII) that can aid in identification, with embeddings from speaker recognition algorithms containing the most information, followed by phone embeddings, linguistic embeddings, and fundamental frequency. Our initial experiments show that the disclosure of PII increases with the length of test samples, but it is bounded by the length of database templates. The provided metric, similarity rank disclosure, provides a way to compare the disclosure of PII between biometric features and merge them to aid identification. It can thus aid in the holistic evaluation of threats to privacy in speech and other biometric technologies.**

Index Terms: privacy-preserving speech processing, order statistics, privacy disclosure, streaming data

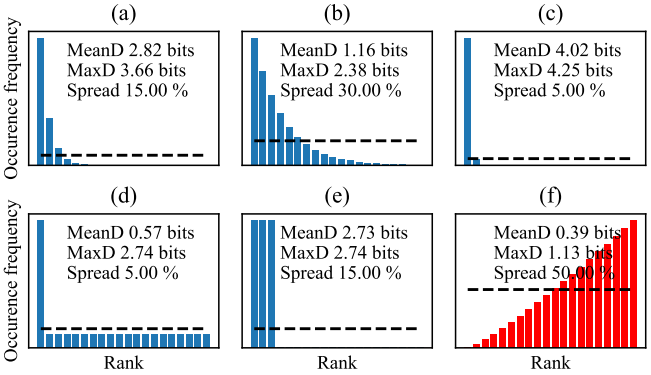

Fig. 1. Illustrations of similarity rank distribution (a) with a typical distribution, (b) with a wide distribution corresponding to noisy data or a poor similarity metric, (c) a narrow distribution corresponding to a good similarity metric and data, (d) where non-matching templates have equal probability, (e) where the most similar templates have equal probability and (f) an anomalous distribution where higher ranks have higher probability. The black dashed line indicates the level of equal probability for all ranks. The mean and maximum disclosures (MeanD and MaxD, respectively) and spread of ranks is listed in each pane (see section IV-C).

I. Introduction

Speech is a tool for communication, conveying private information from one agent to another. In addition to the legitimate message, speech contains plenty of bundled side information, such as state of health, emotions, identity, and background, most of which is private [1, 2]. Speech technology, which processes, transmits, or stores speech, can thus expose users to a multitude of threats like price gouging, stalking, and identity theft [3]. Protecting user privacy in speech applications is morally important, but since threats to users also reduce user satisfaction, it is also important for business.

To accurately characterize privacy threats, we need to quantify them. In speech technology, the predominant approach for identifying speakers is to compare an input signal to a database of speaker templates, and the speaker in the database most similar to the input is chosen as the identified user. However, if the database is large or if the speaker identification approach is noisy, we are unlikely to identify the correct speaker. Still, if a particular speaker template is a "relatively good" match to the observation, it does support the hypothesis that this template is the true identity. In other words, even when the true speaker is not the most similar in the database, observations can still disclose an amount of private information.

Categories of speakers' properties that can be used to identify them are plentiful and include at least physical and psychological characteristics and state, linguistic content and style, language used, proficiency and accent, affiliations, relationship character, acoustic environment, as well as hardware and software used for recording [1, 3]. Information in every such category can potentially aid in identifying a speaker, in particular if an attacker has access to multiple categories of information [2]. It is thus essential to quantify privacy in a way that enables the quantification of privacy from combinations of categories.

The purpose of this paper is to quantify the privacy disclosure of rank-order data from a similarity measure in terms of the reduction in entropy obtained through an observation. It applies to scenarios where similarity measures are imperfect

Bäckström, Nguyen and Rech are with Aalto University, Department of Information and Communications Engineering, Finland, email:first.lastname@aalto.fi. Vali is with Aalto University, Department of Computer Science, Finland.

Manuscript received April 19, 2021; revised August 16, 2021.

A part of this work was supported by "Designing Inclusive & Trustworthy Digital Public Services for Migrants in Finland (Trust-M)" (Grant Number: 353529) funded by the Strategic Research Council of Finland.

0000–0000/00$00.00 © 2021 IEEE

or where data is noisy (see section III). We provide a methodology for quantifying such disclosure from histograms of similarity rank extracted from large databases (see section IV). We also derive a model that can be applied to small or sparse datasets, based on the beta-binomial distribution. We apply the methodology and model to histograms of fundamental frequencies and phone distributions, speaker embeddings from a speaker recognition algorithm, and linguistic embeddings (see section VI) using the Fisher and VoxCeleb corpuses (see section V). Our experiments in section VII demonstrate that the beta-binomial distribution is an accurate model of empirical distributions of similarity rank. Additionally, we find that speech attributes carry identifying information (in an increasing order) in fundamental frequency distribution, linguistic style, phone distribution, and speaker identity embeddings. The magnitude of disclosure depends on both the length of the observation (test sample) and the template database entries. The disclosure, however, saturates after a certain length of the test/template, depending on the length of the other.

II. Background

The predominant approach for privacy-preserving processing of speech signals is *disentanglement*, where the signal is decomposed into components representing separate features of the speech signal, such as its phonetic content, intonation, and speaker identity [3–6]. Provided that the disentangled channels represent independent information, they can be selectively anonymized through replacement, distortion, or quantization [4, 7]. Replacement of speaker characteristics is known as *voice conversion*, and when the purpose of such conversion is to protect privacy, it is known as *anonymization*.

Though disentanglement is an essential element of privacy-preserving speech processing, it is crucial to understand where private information resides and who has access to it. A recommended way of studying, characterizing, and encapsulating such information is a scenario of use or threat model [3, 8]. In particular, in scenarios involving interactions between local (edge) devices and cloud servers, it may be beneficial to protect privacy as early in the pipeline as possible, or in practice, on the edge device [3, 9]. This limits the opportunities that stakeholders and attackers have to abuse private information.

A notable activity in privacy-preserving speech research is the semi-regular VoicePrivacy Challenge [10], where privacy-preserving processing methods are evaluated. The natural counterpart to privacy-preserving speech processing is speaker recognition [11, 12], where the objective is to identify *who* is speaking. Typically, both privacy-preserving algorithms and speaker recognition are evaluated with metrics related to the Equal Error Rate (EER), where the detection threshold is adjusted such that false acceptance rate (FAR) and false rejection rate (FRR) are equal [10, 13]. These measures, however, have significant drawbacks: they quantify only the mean exposure, whereas worst-case exposure should also be considered [14]; they employ pairwise comparison without an explicit joint probability model, and they do not address the magnitude of disclosure for private information.

A significant proportion of research on privacy is related to databases, where fixed-length data is communicated between agents. This work, in contrast, applies also to streaming data, whose impact on privacy is by nature more difficult to analyze [15]. A series of queries to the same database share similarities with streamed data and thus the composition and advanced composition theorems apply, according to which privacy disclosure scales with length T, respectively, as $\mathcal{O}(T)$ and $\mathcal{O}\left(\sqrt{T}\right)$ [16]. However, in streaming, subsequent data points are typically neighbors, whereas subsequent queries are typically random points [15]. We can expect that this influences the disclosure of private information, though it is not immediately clear whether it increases or decreases the threat. In any case, a proposal for Principles of Hippocratic Stream Data Management Systems [15] is for privacy-preserving processing of streamed data to ensure 1) Purpose Specification, 2) Consent, 3) Limited Collection, 4) Limited Use, 5) Limited Disclosure, 6) Limited Retention, 7) Accuracy (in terms of sampling, quantization, and delay), 8) Safety, 9) Openness (for the data owner to access information about themself), 10) Compliance (for the data owner to verify compliance with privacy-preserving principles), 11) Faithful Representation (of the underlying, accurate data), and 12) Minimum Quality of Service (QoS). The current work focuses on principle 7 by quantifying disclosure in streamed systems, as well as evaluation of the limitations of purposes 3 to 6.

The statistics of rank-order data, such as that discussed in this work, may be studied using *order statistics* (e.g. [17, Section 5.5]). Such tools can be used to determine, for example, the probability that an input is the nth largest in a set, or the probability distribution of the same. More specifically, assessing the accuracy of biometric identification has been characterized with the beta-binomial distribution [18]. However, as far as the authors know, the beta-binomial distribution has not been applied to real-world data, and speech applications in particular, and its application has mainly been from a biometric identification rather than a privacy-preserving perspective [19, 20]. A crucial difference in the two approaches is that biometric identification is concerned with the probability of correct identification (and its confidence intervals). In contrast, in privacy-preserving processing, we are concerned with the amount of information revealed about a user (e.g., in bits). Notably, the recognition probability depends on the size of the template database, whereas information disclosure should be independent of that.

The theoretical analysis of privacy has become almost synonymous with *differential privacy* [21]. It is a mathematical framework for ensuring that the analysis of a dataset does not disclose information about individual data points, such as individual users. A central tool for ensuring such protection is the addition of noise to the outputs (or intermediate points) of the protected system. While the framework gives strong tools for ensuring privacy, it has been adopted in only a few speech processing applications [22–25].

Metrics for evaluating privacy-preserving speech processing technologies include 1) the *Equal Error Rate (EER)*, which is defined as the threshold where the rates of False Positives and False Negatives are equal [26, 27], 2) the Privacy *ZEBRA* metric [14], which assesses the average and worst-case

TABLE I
SCENARIO OF USE FOR THE PRIVACY-PRESERVING APPLICATION FOLLOWING [8].

Attacker Model		**Protector Model**	
Objective	Reidentify user from anonymized sample.	Objective	**Defense objective:** Protect identity of user. **Utility objective:** Retain the utility of the downstream application.
Opportunity	Access to anonymized sample (public data).	Opportunity	An anonymization method is available.
Resources	Database of biometric templates (found data).	Resources	None.

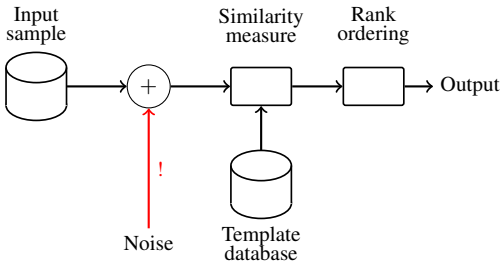

Fig. 2. The *threat model* for evaluating the accuracy of biometric recognition, where the input is corrupted by noise, thwarting the accuracy of the similarity measure to templates in the database and thus altering their rank ordering. The unfavorable insertion of noise is highlighted with a red line and an exclamation mark.

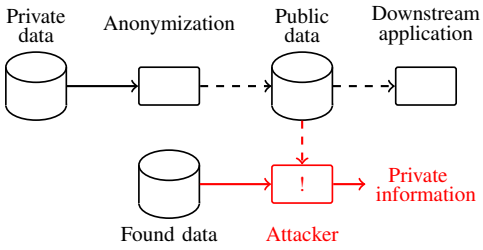

Fig. 3. The *threat model* for evaluating privacy-preserving anonymization following [27], where private data is anonymized to remove private information, and the anonymized data is shared publicly. An attacker uses any available (found) data and anonymized public data to infer private information contrary to the users' preferences. The anonymized data flow is indicated by dashed lines, and the attack by red lines and an exclamation mark.

protection of a privacy-protection system, 3) the *application-independent log-likelihood-ratio cost function* $C_{llr}^{\min}$ that generalizes the EER by considering optimal thresholds over all possible prior probabilities and all possible error cost functions [27, 28], 4) the *maximum achievable key length from the Receiver Operating Characteristic* (ROC) of an optimal biometric recognition system [29], 5) the *tandem equal error rate* (t-EER) that jointly evaluates countermeasures and biometric comparators [30], 6) *linkability* $D_{\leftrightarrow}^{sys}$ computes the overlap in scores between mated (same speaker) and non–mated (different speakers) scores [31], and 7) k*-anonymity*, where a speaker cannot be distinguished within a group of k speakers [32]. Out of these, only Metrics 2 and 4 provide disclosure in terms of entropy (bits) and are thus compatible with the theory of differential privacy. Similarly, only Metric 6 explicitly considers the distribution of *other* speakers. None of them combines these two beneficial properties.

III. THREAT MODELS

The threat model of biometric identification we use is illustrated in fig. 2. We assume that all modules of the identification application are functioning correctly; however, the input sample is noisy, which reduces the accuracy of the similarity measure and corrupts the rank order. This is functionally equivalent to a noise-free input but with an unreliable similarity measure that may not always return a correct answer.

The threat model of the privacy-preserving applications is illustrated in fig. 3, based on [27]. The attacker and protector's objectives, opportunities, and resources are further characterized in a scenario of use in table I, following [8]. We assume that private data, including biometric information, is anonymized before it becomes public. Thus, the attacker has reduced opportunities to re-identify the input sample. However, the attacker can gain aid from access to "found data", corresponding to any information source. Observe that from the attacker's perspective, fig. 3 is identical to fig. 2, as "found data" and "anonymization" in fig. 3 correspond to the "template database" and "noise" in fig. 2. The attacker's objective is the opposite of the user's. The two threat models are thus, for our purposes, otherwise identical except that the optimization objectives have opposite signs. For the quantification of private information, this sign does not make a difference, and we can use fig. 2 as our threat model without reducing generality.

IV. SIMILARITY RANK DISCLOSURE

Suppose we have a dataset of speaker templates, an input speech signal and a similarity metric. Assuming one of the templates correspond to the input speech signal, we can calculate the similarity between all templates and the input. If the data and similarity metrics are accurate, then the most similar template should correspond to the true speaker. However, when the similarity metric or data is imperfect, and especially when speakers have similar voices, the most similar template may not match the true speaker. We can thus rank order the templates according to the their similarity metrics. Similar voices will then a low rank and dissimilar voices a high rank.

The location k of the true speaker in the similarity rank is the central quantity. The probability that the true speaker will have similarity rank k is then $p(k)$. Figure 1 illustrates possible similarity rank distributions. Pane (a) corresponds to a typical situation where the lowest (most similar) ranks are most likely. Noisy data and a poor similarity measure will give a wide distribution (pane (b)). A good similarity metric and data will give a narrow distribution centered at the low end

(pane (c)). The shape of the distribution also bears a message of the setup. At one extreme case, if the non-matching ranks (rank > 1) are equally likely (pane (d)), it means that a non-matching observations bear a minimum of information about the true speaker. The other extreme is when k lowest ranks have equal probability (pane (e)), giving k-anonymity for the speaker [32]. The last case, pane (f), illustrates a similarity rank histogram where higher ranks have higher probability. This anomalous case indicates that the similarity rank is working incorrectly, since dissimilarity is a better indication of identity than similarity.

A. Modelling Rank Distribution

Suppose we have a database of user templates, $c_k \in \mathbb{R}^{M\times 1}$, $k \in [1, N]$, and an input $x \in \mathbb{R}^{M\times 1}$. We can calculate the distance between the input and each of the templates as $d_k = d(c_k, x)$. Suppose c_l, for some $l \in [1, N]$ matches the true identity of the input. Let the probability that the distance to the true identity d_l is smaller than the distance to some other template d_k be

$$P[d_l < d_k] = \theta_k. \tag{1}$$

The probability that d_l is the smallest of all N templates is then

$$P\left[\cup_{k\neq l} d_k < d_l\right] = \prod_{k\neq l} \theta_k. \tag{2}$$

With a large database size N, this becomes increasingly improbable. We can thus instead consider the probability that c_l is the jth best match.

If the θ_k's are drawn from a beta-distribution with parameters α, β, then the probability that c_l is the jth best match follows the beta-binomial distribution [33]

$$\gamma_j := P\left[j = \arg\min_k d_k\right] = \binom{N}{j} \frac{B(j+\alpha, N-j+\alpha)}{B(\alpha,\beta)}, \tag{3}$$

where $B(\cdot)$ is the beta function

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{4}$$

From the perspective of an attacker for whom the true identity is not known, eq. (3) gives the posterior probability that the true identity is j. Assuming all identities were equally likely before the observation, the prior probability is $\frac{1}{N}$. The disclosure of private information or the *rank order disclosure* is thus

$$\epsilon_j := \log_2 N - \log_2 \gamma_j. \tag{5}$$

Note that this disclosure provides information about all identities in the database, regardless of whether they match the input or not. However, the disclosure may or may not support identification. A positive disclosure, $\epsilon_j > 0$, means that the probability that the input corresponds to speaker j is larger than before the observation. Correspondingly, a negative disclosure, $\epsilon_j < 0$, means that the probability that the observation is speaker j has diminished.

B. Parameter Estimation

To use eq. (3), we need to estimate the parameters α, β from data. Our informal experiments have shown that the trivial method of moments approach generally yields nonsensical values (negative α, β) [33]. With a histogram of observations, h_k, $k \in [1, N]$, normalized to the empirical probability distribution $\tilde{p}_k := \frac{h_k}{\sum_k h_k}$, we optimize γ_k using the objective functions:

log-likelihood (LL)	$\min_{\alpha,\beta} -\sum_k \tilde{p}_k \log \gamma_k$
mean square (MS)	$\min_{\alpha,\beta} \sum_k (\tilde{p}_k - \gamma_k)^2$
weighted mean square (WMS)	$\min_{\alpha,\beta} \sum_k \tilde{p}_k (\tilde{p}_k - \gamma_k)^2$
rank-weighted mean square (RWMS)	$\min_{\alpha,\beta} \sum_k e^{-k} (\tilde{p}_k - \gamma_k)^2$
constrained log-likelihood (CLL)	$\min_{\alpha,\beta} -\sum_k \tilde{p}_k \log \gamma_k + 10^5 (\tilde{p}_1 - \gamma_1)^2$.

The goodness-of-fit for the estimated models are evaluated with two approaches; the base-2 Kullback-Leibler (KL) divergence and $\log_2$-match for the first rank:

KL-divergence	$\sum_k \tilde{p}_k \log_2 \frac{\tilde{p}_k}{\gamma_k}$
rank-1 match	$\left\lvert \log_2 \frac{\tilde{p}_1}{\gamma_1} \right\rvert$

where $\tilde{p}_k$ and γ_k are the normalized histogram (empirical probability distribution) and beta-binomial model, respectively. The KL-divergence thus quantifies overall performance, while the rank-1 match quantifies goodness-of-fit for the most important rank, i.e. $\mathrm{rank} = 1$.

C. Rank Disclosure Statistics

Beyond the disclosure of individual observations, it is essential to characterize the statistical properties of disclosure. In particular, both the mean and worst-case disclosure are of interest [7, 14]. With k as the rank of the true speaker among N speaker and with the empirical probability distribution $\tilde{p}_k$ (or its corresponding model γ_k) the pertinent statistics are:

individual disclosure	$\epsilon_k := \log_2 N - \log_2 \tilde{p}_k$
mean disclosure (MeanD)	$\mu := \sum_k \tilde{p}_k \epsilon_k$
identification rate (IdR)	$\tilde{p}_1$
standard deviation of mean disclosure (StDD)	$\sigma := \sqrt{\sum_k \tilde{p}_k (\epsilon_k - \mu)^2}$
maximum disclosure (MaxD)	$\max_k \epsilon_k$
rank spread (Spread)	$\frac{1}{N}\mathrm{count}_{k\in[1,N]} \left(\tilde{p}_k > \frac{1}{N}\right)$.

Here, the *rank spread* encompasses the proportion of bins where the (empirical) probability is above the equal-probability level (i.e. pure chance), using the $\mathrm{count}_k()$ function that counts the number of true expressions.

Figure 1 illustrates the mean disclosure (MeanD) and rank spread of sample histograms.

D. Summary

The proposed metric quantifies the disclosure of personally identifying information (PII) from similarity rank information. For a similarity measure $S(x, y)$, the required steps are

1) For every test input x_k
   a) Evaluate the similarity $S(x_k, y_h)$ to every database template y_h.
   b) Rank order similarity scores such that the most similar is first.
   c) Store the rank of the true match in the histogram.
2) Normalize histogram to obtain the empirical probability distribution.
3) If the distribution is sufficiently dense (plenty of data), continue to step 5.
4) Fit a beta-binomial distribution to the empirical probability distribution (see section IV-B).
5) Evaluate the disclosure statistics for the histogram or beta-binomial model (see section IV-C).

V. Data

We use three speech corpora for our experiments, the Fisher English Training Speech Part 2 corpus [34], VoxCeleb [35], and a new purpose-built dataset, LibriLong, based on the LibriVox free public domain audiobooks[1], similar to LibriSpeech [36]. The Fisher dataset contains 975 h of spontaneous, conversational speech at a sampling rate of 8 kHz. There are 5848 conversations of 10 min each, with the two speakers in separate channels, which gives a total of 11 696 unique speakers. VoxCeleb contains 1251 speakers, each with an average of 116 utterances, each of length 8.2 s on average. Speaker identities in Fisher and VoxCeleb thus correspond to 13.5 bit and 10.3 bit of information, respectively.

The LibriLong dataset contains 64 excerpts of audiobooks in English, each of at least 2 h in length, sampled at 16 kHz. Speaker identities in LibriLong thus contain 6 bit of information. The purpose of the dataset is to allow evaluation of privacy threats in long recordings. Excerpts were randomly chosen from the "general fiction" category of LibriVox, such that a single person read the whole excerpt, and the excerpt is of at least 2 h in length. The distribution of perceived genders of speakers (as judged by the current authors) is 31 male and 33 female. One candidate was replaced because the fundamental frequency distribution was informally found to differ significantly between chapters of the audiobook. This dataset is published openly[2]. To allow comparison between the Fisher and LibriLong datasets, in our experiments, LibriLong was resampled to 8 kHz. LibriLong is read text, and the linguistic content thus relates to the author of the audiobook rather than the speaker. Consequently, we did not use LibriLong for linguistic analysis.

VI. Speaker features

We characterize speakers with four different features, 1) the fundamental frequency distribution, 2) a phone embedding distribution, 3) a speaker identity embedding, and 4) a text style embedding. This provides a range of abstraction levels for the personally identifying information (PII) of speakers. Notably, the aim is not to develop new or improved features, but we chose standard approaches to demonstrate the use of the proposed disclosure statistics.

The fundamental frequency of a speech sample is analyzed with the YIN algorithm [37] using the Torch-YIN implementation[3]. We limit the frequency range of the fundamental to 65 to 450 Hz. This implementation estimates pitch lags as integers, which gives 107 distinct fundamental frequencies. The stride is 10 ms such that each 30 s segment contains up to approximately 3000 measurements, but since non-voiced speech frames are discarded, the actual number of measurements per segment is lower. The histogram of fundamental frequencies over the entire Fisher dataset is illustrated in fig. 4. The histograms h_k are normalized to the corresponding empirical probability distribution, $\tilde{p}_k = \frac{h_k}{\sum_k h_k}$.

Phone embeddings are extracted from a VQ-VAE-based voice conversion algorithm [38, 39]. We trained the encoder, decoder, and space-filling vector quantizer of [39] on the *ZeroSpeech 2019 Challenge* English dataset [40] with the same hyper-parameters as in [39], while the speech files were resampled to 8 kHz to train the model on a similar sampling rate as the Fisher dataset. The quantization of the phone embeddings is done over 256 codebook vectors (i.e., 256 bins). Each 10 min conversation of Fisher is split into twenty 30 s chunks. For each chunk and each speaker, the silence and the frames shorter than 0.1 s speech are skipped for phone extraction. The histogram of phone embeddings over the entire Fisher dataset is illustrated in fig. 4. Again, the histograms h_k are normalized to the corresponding empirical probability distribution, $\tilde{p}_k = \frac{h_k}{\sum_k h_k}$.

Speaker identity embeddings are extracted from each audio file in VoxCeleb using the standard baseline for speaker recognition, ECAPA-TDNN [41], using its pre-trained SpeechBrain implementation [42]. To get an equal number of observations for each speaker, we use the 45 first files of each speaker as the observations. The average EER for this dataset was 0.76 %, which is low enough to confirm that ECAPA-TDNN works properly for this dataset.

The linguistic features are quantified with Linguistic Inquiry and Word Count (LIWC). This computerized text analysis method is widely used to analyze emotional, cognitive, and structural components in written and spoken language [43, 44]. We use all linguistic features provided by the LIWC-22 dictionary, except those specifically measuring the text content domain, as the conversation topics were not freely chosen in the Fisher dataset. Since our data consists of transcripts, we also consider conversational features such as disfluencies and the frequency of filler words. Combining these elements, we construct an 80-dimensional feature vector. We split the text of the Fisher dataset of each speaker into 5 segments of 2 min. Utterances were assigned to segments according to the location of segment centers, $T_{center} = \frac{1}{2}(T_{start} + T_{end})$.

[1] https://librivox.org/

[2] URL published when paper is accepted

[3] https://github.com/brentspell/torch-yin

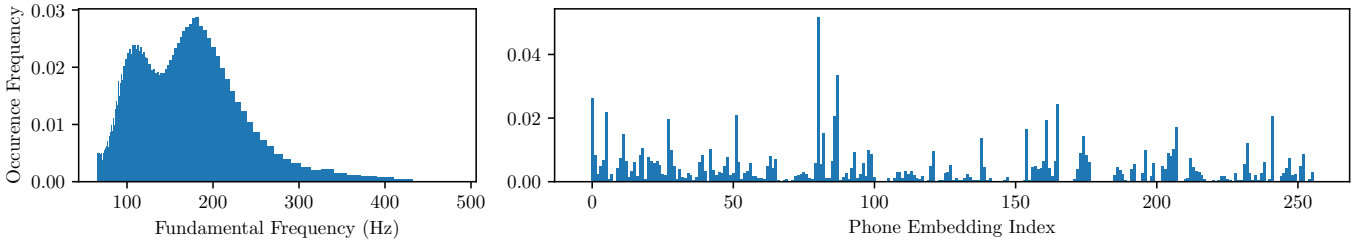

Fig. 4. Histograms of fundamental frequencies and phone embeddings of all speakers in the Fisher dataset.

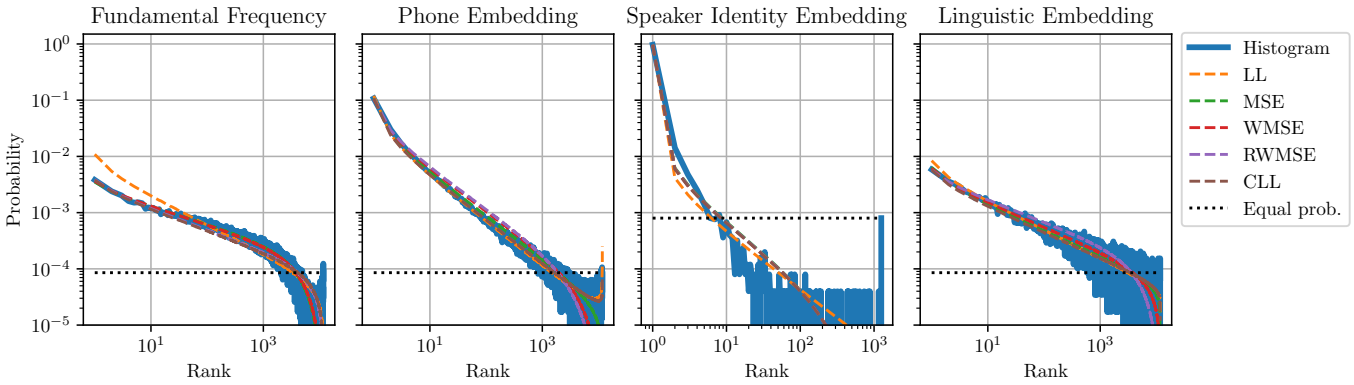

Fig. 5. The empirical probability distribution of similarity rank. The dashed lines indicate beta-binomial models fitted to the data, and the dotted black line indicates the level of equal probability.

VII. Experiments

The purpose of our experiments is to 1) demonstrate the usefulness of the proposed disclosure measures by applying them to different categories of personally identifiable information (PII), 2) get a first impression of the magnitude of PII disclosure in those categories, 3) evaluate the applicability of the proposed beta-binomial model to both dense and sparse histograms, and 4) study disclosure in streamed applications.

We construct experiments as follows: Let the number of speakers be N, the number of observations per speaker be K, and each observation has M dimensions. Let T, D be the number of observations of input samples and the database templates, respectively, and the starting point for input samples be S. The indices for samples are then defined as

$$\begin{array}{ll} \text{input samples} & S + [0 \dots T) \mod K \\ \text{database templates} & S + T + \Delta + [0 \dots D) \mod K, \end{array} \tag{6}$$

where $\Delta = \frac{1}{2}(K - S - T)$. This gives non-overlapping segments for the input and database with maximum separation. By iterating S over $[0, K)$, we obtain data similar to an S-fold cross-validation experiment.

Figure 5 illustrates the similarity rank histograms of the fundamental frequency, phone, speaker, and linguistic embeddings with $T = 1, D = 1$. The histograms are plotted in a log-log graph to focus attention on the lowest ranks and to visualize probabilities of different orders of magnitude. We observe that all four plots have a similar shape; the histograms are more or less linear in the log-log domain, except for the lowest ranks, which bend upward. We can furthermore see that the speaker identity embedding has the highest probability for a match (rank $= 1$), followed by phone and linguistic embeddings, as well as the fundamental frequency. Observe that this comparison should be taken as indicative only, since the speaker identity ranks are calculated from a corpus with a smaller number of speakers, and the numbers are thus not directly comparable.

The disclosure statistics for each case are presented in table II. The mean disclosure and identification rates confirm the observation from fig. 5 that speaker identity embeddings contain the most accurate information. Interestingly, the maximum disclosure is higher for the phone than the speaker embedding. However, since VoxCeleb contains only 10.3 bit of speaker identity information, that will limit the disclosure. We can thus expect that the maximum disclosure from the speaker identity embedding is higher than that. The spread for the disclosure from the fundamental frequency and linguistic embeddings is approximately 28 %. This indicates the proportion of speakers whose recognition probability will, on average, increase for every observation. Observe that these are predominantly false positives, where the posterior probability increases for speakers similar to the test subject. The spread for phone embeddings is significantly smaller, and even smaller for the speaker identity embedding.

The Equal Error Rates (EER) of each feature are largely inversely proportional to the mean and maximum disclosure as well as identification rate. The only apparent deviation is the phone embedding, where the EER is higher than for the fundamental frequency; yet, the disclosures and identification rate are also higher.

To each of the empirical probability distributions (i.e., normalized histograms), we fitted the beta-binomial distribution using the loss functions listed in section IV-B. The models

TABLE II
DISCLOSURE STATISTICS OF SIMILARITY RANK HISTOGRAMS, BETA-BINOMIAL MODEL OPTIMIZED WITH CLL AND THE BASELINE METRIC EQUAL ERROR RATE (EER) CORRESPONDING TO THE FUNDAMENTAL FREQUENCY (F0), PHONE, SPEAKER IDENTITY, AND LINGUISTIC EMBEDDINGS.

Histograms

	F0	Phone	Identity	Linguistic
MeanD (bits)	0.99	2.45	9.97	0.66
IdR (%)	0.38	10.77	96.74	0.58
MaxD (bits)	5.48	10.30	10.24	6.10
StDD (bits)	1.57	3.87	1.57	1.44
Spread (%)	28.15	13.47	0.48	28.22

Model optimized with CLL

	F0	Phone	Identity	Linguistic
MeanD (bits)	0.59	2.27	9.89	0.41
IdR (%)	0.38	10.67	96.75	0.57
MaxD (bits)	5.47	10.29	10.24	6.07
StDD (bits)	1.31	3.79	1.99	1.29
Spread (%)	34.95	14.00	0.64	28.77

Baseline metric

EER (%)	32.30	37.78	0.76	39.10

are illustrated in fig. 5. We observe that though the log-likelihood is the standard approach to fitting distributions, in most cases, it provides a poor fit for the lowest ranks. In the current application, the lowest ranks are the most important, as there the disclosure is the highest. This warranted informal experimenting with different losses (see section IV-B), the most promising of which are included in this paper.

The goodness-of-fit measures of each feature and loss are listed in table III. We observe that the log-likelihood (LL) obtains the best (smallest) KL-divergence in all cases except one, where it is very close. This is natural since the log-likelihood and KL-divergence differ only by a constant scalar offset. The constrained log-likelihood (CLL), which penalizes for mismatch in the rank-1 value, also obtains the best (smallest) rank-1 match, except in two cases where it is very close. As the constrained log-likelihood (CLL) is reasonably good also concerning the KL-divergence, we conclude that it is the best loss for our purposes.

The disclosure statistics for the beta-binomial model in the four cases are listed in table II. Overall, the statistics roughly align with those extracted from the raw histograms. The identification rates (IdR) and maximum disclosure (MaxD) are all match within 1 % difference. The mean disclosures (MeanD) have at most a 0.4 bit difference. Observe that such difference can be expected as the model is an approximation. However, we cannot unequivocally say that either the raw data or the model would be better or more accurate than the other, since data is noisy and the model is a simplification. Still, the differences in disclosures give an indication of the accuracy that can be expected of these metrics. In other words, with the current measurement setup, the mean disclosure metric can have an error of the order of 0.5 bit.

The above histograms were accumulated over all speakers. Individual speakers may, however, have different characteristics, making them easier or harder to identify. We may thus expect that the similarity rank histograms are significantly

TABLE III
GOODNESS-OF-FIT MEASURES FOR THE BETA-BINOMIAL MODEL OPTIMIZED WITH DIFFERENT LOSS FUNCTIONS (SEE SECTION IV-B) AND FOR EACH OF THE SPEECH FEATURES. SMALLER IS BETTER. THE BEST LOSS FOR EACH COMBINATION OF FEATURE AND METRIC IS IN BOLD.

Metric	Loss	F0	Phone	Identity	Linguistic
KL-div. (bits)	LL	$\mathbf{1.66\times10^{-1}}$	4.97×10^{-2}	$\mathbf{3.94\times10^{-2}}$	$\mathbf{1.57\times10^{-1}}$
	MSE	3.94×10^{-1}	2.12×10^{-1}	6.70×10^{-2}	1.66×10^{-1}
	WMSE	3.50×10^{-1}	7.97×10^{-1}	6.53×10^{-2}	2.64×10^{-1}
	RWMSE	1.97×10^{-1}	1.25	6.65×10^{-2}	6.78×10^{-1}
	CLL	1.96×10^{-1}	$\mathbf{4.87\times10^{-2}}$	6.54×10^{-2}	1.70×10^{-1}
Rank-1 match (bits)	LL	1.52	1.45×10^{-1}	1.11×10^{-3}	5.28×10^{-1}
	MSE	1.33×10^{-1}	1.32×10^{-2}	1.40×10^{-3}	1.37×10^{-1}
	WMSE	7.72×10^{-2}	4.50×10^{-2}	$\mathbf{1.92\times10^{-4}}$	6.92×10^{-2}
	RWMSE	2.98×10^{-2}	1.25×10^{-2}	3.46×10^{-4}	$\mathbf{1.66\times10^{-2}}$
	CLL	$\mathbf{1.63\times10^{-3}}$	$\mathbf{7.94\times10^{-3}}$	1.03×10^{-3}	2.65×10^{-2}

different for individual speakers. Figure 7 illustrates the histogram and the corresponding empirical cumulative distribution (scaled to unity) of an example speaker. We use linear axes since the sparse histogram is mostly zero, which cannot be displayed in a log-log plot. The histogram is also always only one or zero, such that the empirical cumulative probability distribution serves as a more informative visualization. We included here only models fitted with LL and CLL, since the other models did not always converge.

We observe that the model optimized with the log-likelihood (LL) criteria visually follows the cumulative distribution more closely than the constrained alternative (CLL). A practical issue is that, in many cases, such as in this plot, the histogram value at rank 1 is zero, although we can reasonably expect it to be positive. This means that optimizing CLL will try to force the model to zero at rank-1, even though that contradicts our expectations of the true probability distribution. This introduces a random component to models optimized with CLL. This leaves the log-likelihood loss as the only remaining viable option.

Figure 6 illustrates the empirical probability distribution of mean disclosure of the beta-binomial models fitted to the individual speakers' histograms. This can be interpreted as the probability distribution of the privacy leak associated with each feature. The maximum disclosure is close to the worst-case for all features, matching the maximum disclosure of the corresponding datasets, which indicates that some speakers are outliers and can be perfectly identifiable from any of the features. This is a concern, as the average disclosures are, in all cases except the speaker identity embedding, reasonably low, in the range 0.9 bit to 2.4 bit. In other words, a low mean disclosure does not automatically indicate that all individuals would be protected from identification. We can, however, see that the distributions of the fundamental frequency, the phone, and linguistic embeddings have a distinct peak around which a majority of individuals are focused. The standard deviations are also reasonably small, below 2 bit. This indicates that a majority of speakers have a similar mean disclosure, and the worst-case scenario concerns only a minority of speakers. The speaker identity embedding peaks at the edge of the range of possible disclosures, indicating that a large proportion of speakers can be perfectly identified. We hypothesize that if

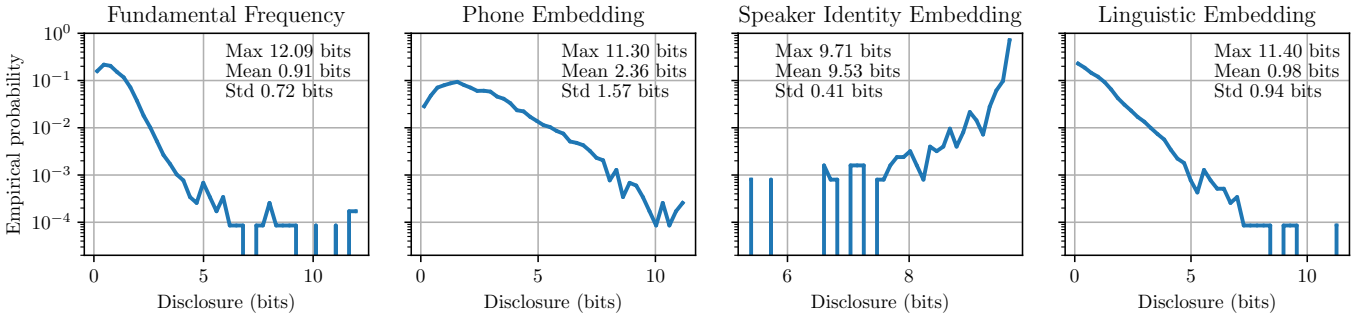

Fig. 6. The empirical probability distribution of the mean disclosure for each feature, calculated from the beta-binomial models fitted with the log-likelihood criteria to the similarity rank histograms. The maximum, mean and standard deviation (std) of the mean disclosures are included for each feature.

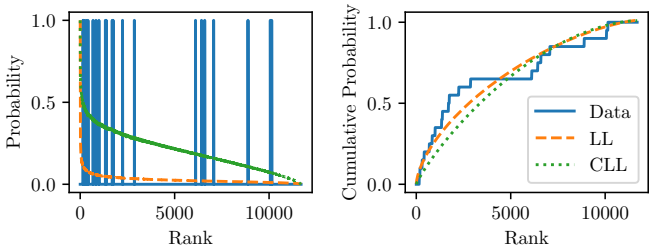

Fig. 7. The histogram/probability distribution (left pane) and the empirical cumulative probability distribution (right pane) for the similarity rank obtained from the fundamental frequency of one speaker in the Fisher dataset, as well as the corresponding beta-binomial models fitted with two different losses, LL and CLL. The probability distributions of the models are scaled to unity for improved visibility.

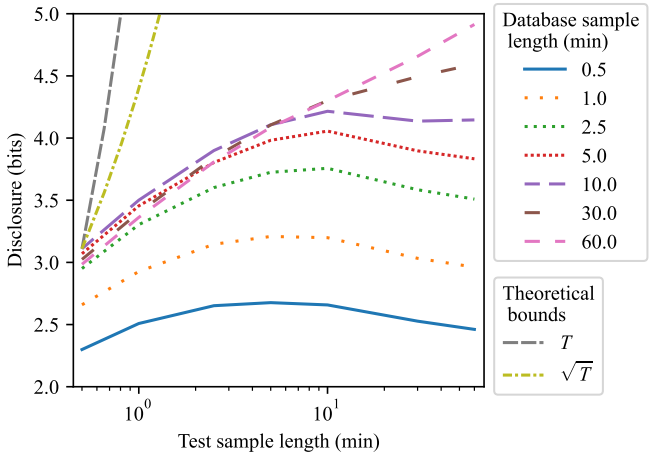

Fig. 8. Disclosure of the fundamental frequency as a function of length of segments used for database templates and test samples. The dashed lines show the theoretical limits to scaling of disclosure based on the composition (T) and advanced composition ($\sqrt{T}$) theorems [16].

the set of speakers and correspondingly, the range of possible mean disclosures were larger, then the speaker identity embedding would also have a distinct peak. Then all four cases would show a similar shape of the distribution.

Finally, fig. 8 illustrates how disclosure of the fundamental frequency scales with the length of the database samples and test samples. We observe that disclosure initially increases with the length of the database samples and test samples, but counterintuitively, it diminishes slightly at higher lengths. Only database samples of lengths 30 and 60 min depart from this pattern, but we assume that with longer test samples, the pattern would repeat. We hypothesize that the increase in accuracy with longer observations causes the initial increase in disclosure. However, speech features are not stationary due to intra-speaker variance, which causes a bias between the database template and test samples that becomes more pronounced in longer observations. This bias is the likely cause for the reduction in disclosure at higher test sample lengths. Due to symmetry between the test and database samples, the same observations hold with their roles reversed.

According to the basic and advanced composition theorems, disclosure scales with length T, respectively, as $\mathcal{O}(T)$ and $\mathcal{O}\left(\sqrt{T}\right)$ [16], illustrated by the long dashed and dashed-dotted lines. We observe that the measured disclosure scales orders of magnitude slower than the theoretical bounds.

VIII. Discussion

Quantification is central to evaluating privacy threats. We propose a new metric, named *similarity rank disclosure*, that quantifies the amount of personally identifying information (PII) contained in an observation. The metric measures entropy (bits) and thus aligns with the theory of differential privacy.

A central outcome of the proposed metric is that the posterior probability of *every* speaker will depend on the observation. In other words, every observation carries information about every speaker. The posterior probability of speakers similar to the test sample will increase, while that of dissimilar speakers will decrease. This insight carries interesting consequences for privacy. If speakers A and B are similar, and A is found to have property X, then that increases the probability that speaker B has the same property. In other words, *an observation of speaker A can violate the privacy of speaker B, even when speaker B has no interaction with the attacker*. This highlights that privacy is not only about individual users, but it has to be evaluated also on community- and societal levels (cf. [45]).

A second outcome is that the posterior of a speaker can decrease due to an observation, even for the true speaker. A decrease in the posterior means that the disclosure has a negative sign and is thus “negative information”. We can thus categorize observations and their corresponding disclosures as:

	Disclosure	
Speaker	Positive	Negative
Match	True positive	False positive
Other	False positive	True negative

A third outcome is that while privacy disclosure in streaming applications is, in theory, increasing without bounds with time, the situation is not quite that bleak in practice. First, when the observation is arbitrarily long, the limiting factor for disclosure is the size of the database. Second, intra-speaker and recording environment variance will significantly reduce the potential disclosure. This does not mean that streaming would be "safe", but only that the privacy exposure is not as extreme as the theoretical limits would suggest.

In our experiments, we evaluated the disclosure of PII from the fundamental frequency and phone distributions, as well as speaker recognition and linguistic embeddings. These features were selected to represent different levels of abstraction, to demonstrate the properties of the proposed metric. Notably, the analysis of these features was implemented using standard tools without any effort to improve them, to maintain focus on the metric. In particular, it is clear that a great deal more information could be extracted from, for example, the fundamental frequency by analyzing its trajectories over time. Such analysis is left for future studies.

In conclusion, we have demonstrated that the proposed metric can be applied to different speech features. The magnitude of disclosure in each case aligns with the authors' intuitive expectations, in that all features do leak PII, and the speaker recognition embedding, specifically developed to reveal the identity of the speaker, has the highest disclosure. For cases with sparse data, we furthermore proposed a model for the probability distribution. The proposed method is thus applicable to a wide range of speech, text, and other biometric applications that require the evaluation of disclosure of personally identifying information (PII).

REFERENCES

[1] Rita Singh. *Profiling humans from their voice*. Vol. 41. Springer, 2019. URL: https://doi.org/10.1007/978-981-13-8403-5.

[2] Rita Singh and Bhiksha Raj. "Human Voice is Unique". In: (2025). arXiv: 2506.18182 [cs.SD]. URL: https://arxiv.org/abs/2506.18182.

[3] Tom Bäckström. *Privacy in Speech Technology*. 2023. URL: https://doi.org/10.48550/arXiv.2305.05227.

[4] Champion Pierre, Anthony Larcher, and Denis Jouvet. "Are disentangled representations all you need to build speaker anonymization systems?" In: *Proc. Interspeech*. 2022, pp. 2793–2797. URL: https://doi.org/10.21437/Interspeech.2022-10586.

[5] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion". In: *Interspeech 2021*. 2021, pp. 1349–1353. URL: https://doi.org/10.21437/Interspeech.2021-319.

[6] Lifa Sun et al. "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training". In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. 2016, pp. 1–6. URL: https://doi.org/10.1109/ICME.2016.7552917.

[7] Mohammad Hassan Vali and Tom Bäckström. "Privacy PORCUPINE: Anonymization of Speaker Attributes Using Occurrence Normalization for Space-Filling Vector Quantization". In: *Proc. Interspeech*. 2024.

[8] Mehtab Ur Rahman et al. "Scenario of Use Scheme: Threat Modelling for Speaker Privacy Protection in the Medical Domain". In: *4th Symposium on Security and Privacy in Speech Communication*. 2024, pp. 21–25. URL: https://doi.org/10.21437/SPSC.2024-4.

[9] Peter Wu et al. "Understanding the tradeoffs in client-side privacy for downstream speech tasks". In: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2021, pp. 841–848. URL: https://doi.org/10.48550/arXiv.2101.08919.

[10] Michele Panariello et al. "The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 3477–3491. URL: https://doi.org/10.1109/TASLP.2024.3430530.

[11] Zhongxin Bai and Xiao-Lei Zhang. "Speaker recognition based on deep learning: An overview". In: *Neural Networks* 140 (2021), pp. 65–99. ISSN: 0893-6080. URL: https://doi.org/10.1016/j.neunet.2021.03.004.

[12] Nik Vaessen and David A Van Leeuwen. "Fine-tuning wav2vec2 for speaker recognition". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7967–7971. URL: https://doi.org/10.1109/ACCESS.2021.3084299.

[13] Andrew Brown et al. *VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge*. 2022. arXiv: 2201.04583 [cs.SD]. URL: https://arxiv.org/abs/2201.04583.

[14] Andreas Nautsch et al. "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment". In: *Proc. Interspeech*. 2020, pp. 1698–1702. URL: https://doi.org/10.21437/Interspeech.2020-1815.

[15] Mohamed H. Ali, Mohamed Y. ElTabakh, and Elisa Bertino. "Hippocratic data streams-concepts, architectures and issues". In: 05-025 (2005). URL: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2638&context=cstech.

[16] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. "The composition theorem for differential privacy". In: *International conference on machine learning*. PMLR. 2015, pp. 1376–1385. URL: https://doi.org/10.1109/TIT.2017.2685505.

[17] George Casella and Roger Berger. *Statistical inference*. New York: CRC press, 2024. URL: https://doi.org/10.1201/9781003456285.

[18] Michael E Schuckers. "Using the beta-binomial distribution to assess performance of a biometric identification device". In: *International Journal of Image and Graphics* 3.03 (2003), pp. 523–529. URL: https://doi.org/10.1142/S0219467803001147.

[19] S.C. Dass, Yongfang Zhu, and A.K. Jain. "Validating a Biometric Authentication System: Sample Size Requirements". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12 (2006), pp. 1902–1319. URL: https://doi.org/10.1109/TPAMI.2006.255.

[20] Sinjini Mitra, Marios Savvides, and Anthony Brockwell. "Statistical Performance Evaluation of Biometric Authentication Systems Using Random Effects Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.4 (2007), pp. 517–530. URL: https://doi.org/10.1109/TPAMI.2007.1000.

[21] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. URL: http://dx.doi.org/10.1561/0400000042.

[22] Tiantian Feng, Raghuveer Peri, and Shrikanth Narayanan. *User-Level Differential Privacy against Attribute Inference Attack of Speech Emotion Recognition in Federated Learning*. 2022. URL: https://doi.org/10.48550/ARXIV.2204.02500.

[23] Ali Shahin Shamsabadi et al. *Differentially Private Speaker Anonymization*. 2022. URL: https://doi.org/10.48550/ARXIV.2202.11823.

[24] Tiantian Feng, Raghuveer Peri, and Shrikanth Narayanan. "User-Level Differential Privacy against Attribute Inference Attack of Speech Emotion Recognition on Federated Learning". In: *Proc. Interspeech*. 2022, pp. 5055–5059. URL: https://doi.org/10.21437/Interspeech.2022-10060.

[25] Michael Shoemate et al. *Sotto Voce: Federated Speech Recognition with Differential Privacy Guarantees*. 2022. URL: https://doi.org/10.48550/ARXIV.2207.07816.

[26] Rashid Jahangir et al. "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges". In: *Expert Systems with Applications* 171 (2021), p. 114591. URL: https://doi.org/10.1016/j.eswa.2021.114591.

[27] Mohamed Maouche et al. "A Comparative Study of Speech Anonymization Metrics". In: *Proc. Interspeech*. 2020, pp. 1708–1712. URL: https://doi.org/10.21437/Interspeech.2020-2248.

[28] Niko Brümmer and Johan Du Preez. "Application-independent evaluation of speaker detection". In: *Computer Speech & Language* 20.2-3 (2006), pp. 230–275. URL: https://doi.org/10.1016/j.csl.2005.08.001.

[29] Raymond NJ Veldhuis. "The relation between the secrecy rate of biometric template protection and biometric recognition performance". In: *2015 International Conference on Biometrics (ICB)*. IEEE. 2015, pp. 311–318. URL: http://dx.doi.org/10.1109%2FICB.2015.7139055.

[30] Tomi H Kinnunen et al. "t-EER: Parameter-free tandem evaluation of countermeasures and biometric comparators". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.5 (2023), pp. 2622–2637. URL: https://doi.org/10.1109/TPAMI.2023.3313648.

[31] Marta Gomez-Barrero et al. "General framework to evaluate unlinkability in biometric template protection systems". In: *IEEE Transactions on Information Forensics and Security* 13.6 (2017), pp. 1406–1420. URL: https://doi.org/10.1109/TIFS.2017.2788000.

[32] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570. URL: https://doi.org/10.1142/S0218488502001648.

[33] Rand R Wilcox. "Estimating the parameters of the beta-binomial distribution". In: *Educational and Psychological Measurement* 39.3 (1979), pp. 527–535. URL: https://doi.org/10.1177/001316447903900302.

[34] Christopher Cieri et al. *Fisher English Training Part 2, Speech LDC2005S13. Web Download*. Philadelphia, 2005. URL: https://doi.org/10.35111/dz78-gx84.

[35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. "VoxCeleb: A Large-Scale Speaker Identification Dataset". In: *Interspeech 2017*. 2017, pp. 2616–2620. URL: https://doi.org/10.21437/Interspeech.2017-950.

[36] Vassil Panayotov et al. "Librispeech: an ASR corpus based on public domain audio books". In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2015, pp. 5206–5210. URL: https://doi.org/10.1109/ICASSP.2015.7178964.

[37] Alain De Cheveigné and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930. URL: https://doi.org/10.1121/1.1458024.

[38] Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge". In: *Proc. Interspeech*. 2020, pp. 4836–4840. URL: https://doi.org/10.21437/Interspeech.2020-1693.

[39] Mohammad Hassan Vali and Tom Bäckström. "Interpretable Latent Space Using Space-Filling Curves for Phonetic Analysis in Voice Conversion". In: *Proc. Interspeech*. 2023. URL: https://doi.org/10.21437/Interspeech.2023-1549.

[40] *ZeroSpeech 2019 Challenge Dataset*. [Online; accessed 06.03.2023]. URL: %7Bhttps://download.zerospeech.com%7D.

[41] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". In: *Proc. Interspeech*. 2020, pp. 3830–3834. URL: https://doi.org/10.21437/Interspeech.2020-2650.

[42] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. 2021. arXiv: 2106.04624 [eess.AS]. URL: https://huggingface.co/speechbrain.

[43] Yla R Tausczik and James W Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54. URL: https://doi.org/10.1177/0261927X09351676.

[44] Ryan L Boyd et al. "The development and psychometric properties of LIWC-22". In: *Austin, TX: University of Texas at Austin* 10 (2022), pp. 1–47. URL: https://doi.org/10.13140/RG.2.2.23890.43205.

[45] Tom Bäckström and Fedor Vitiugin. "Beyond User-centric: Modelling Privacy and Fairness Effects of Speech Interfaces on Community- and Society-Levels". In: *accepted to 3rd Symposium on Security and Privacy in Speech Communication*. ISCA, 2025.

Tom Bäckström (Senior Member, IEEE) received the master's and Ph.D. degrees from Aalto University, in 2001 and 2004, respectively, which was then known as the Helsinki University of Technology. He has been an associate professor in the Department of Signal Processing and Acoustics at Aalto University, Finland, since 2016. He was a Professor at the International Audio Laboratory Erlangen, Friedrich-Alexander University, from 2013 to 2016, and a Researcher at Fraunhofer IIS from 2008 to 2013. He has contributed to several international speech and audio coding standards and is the chair and co-founder of the ISCA Special Interest Group "Security and Privacy in Speech Communication". His research interests include technologies for spoken interaction, emphasizing efficiency and privacy, and in particular, in multi-device and multi-user environments.

Mohammad Hassan Vali received his bachelor's and master's degrees from Babol Noshirvani University of Technology in 2014 and 2017, respectively, and his doctorate from Aalto University in 2025. He is currently a postdoctoral researcher in the Department of Computer Science at Aalto University, Finland. His research interests include image and speech processing using machine learning models, with a particular focus on vector quantization in deep neural networks for discrete representation learning.

My Nguyen is a master's student in Speech and Language Technology at Aalto University, Finland. Her research interests include forensic linguistics and ethical AI. She is currently working on a master's thesis related to the application of author attribution methods in speaker identification.

Silas Rech is a doctoral student in the Speech Interaction Technology Group at Aalto University, Finland. He received his master's degree from Aalto in 2022. His research interests focus on trustworthy, private, and ethical speech processing, in particular natural interfaces.